On gaps and unoccupied urns in sequences of geometrically distributed random variables

Guy Louchard* Helmut Prodinger†

May 31, 2007

Abstract

This paper continues the study of gaps in sequences of n geometrically distributed random variables, as started by Hitczenko and Knopfmacher [11], who concentrated on sequences which were gap-free. Now we allow gaps, and count some related parameters.

Our terminology of gaps just means empty "urns" (within the range of occupied urns), if we think about an urn model. This might be called *weak* gaps, as opposed to *maximal* gaps, as in [11]. If one considers only "gap-free" sequences, both notions coincide asymptotically, as $n \to \infty$.

First, the probability $p_n(r)$ that a sequence of length n has a fixed number r of empty urns is studied; this probability is asymptotically given by a constant $p^*(r)$ (depending on r) plus some small oscillations. When $p = q = 1/2$, everything simplifies drastically; there are no oscillations.

Then, the random variable 'number of empty urns' is studied; all moments are evaluated asymptotically. Furthermore, samples that have r empty urns, in particular the random variable 'largest non-empty urn' are studied. All moments of this distribution are evaluated asymptotically.

The behaviour of the quantities obtained in our asymptotic formulæ is also studied for $p \to 0$ resp. $p \to 1$, through a variety of analytic techniques.

The last section discusses the concept called 'super-gap-free.' A sample is super-gap-free, if $r = 0$ and each non-empty urn contains at least 2 items (and d-super-gap-free, if they contain $\geq d$ items). For the instance $p = q = 1/2$, we sketch how the asymptotic probability (apart from small oscillations) that a sample is d-super-gap-free can be computed.

1 Introduction

Let us consider a sequence of n random variables (RV), $Y_1, \dots, Y_n$, distributed (independently) according to the geometric distribution Geom(p). Set $q := 1 - p$, then $\mathbb{P}(Y = j) = pq^{j-1}$. If we neglect the order in which the n items arrive, we can think about an urn model, with urns labelled $1, 2, \dots$, the probability of each ball falling into urn j being given by pq^{j-1}.

*Université Libre de Bruxelles, Département d'Informatique, CP 212, Boulevard du Triomphe, B-1050 Bruxelles, Belgium. `louchard@ulb.ac.be`

†University of Stellenbosch, Mathematics Department, 7602 Stellenbosch, South Africa. `hproding@sun.ac.za`

Set the indicator RV (in the sequel we drop the n-specification to simplify the notations):[1]

$$X_i := [\![\text{value } i \text{ appears among the } n \text{ RVs}]\!],$$

i.e. urn i is not empty.

> *A comment on terminology: In [11], Hitczenko and Knopfmacher consider gap-free distributions, i.e., the indices $a, a+1, \dots, b$ of the non-empty urns form an interval. Without loss of generality one may assume that $a = 1$, since there is only an exponentially small probability for the first urn to be empty. So, the extra notation ("complete") for this instance can be ignored.*
>
> *Gaps themselves are not explicitly mentioned, but it is understood that a gap is a (maximal) sequence of empty urns between non-empty ones.*
>
> *Our point of view is different here: We say that an urn b is a gap, if it is empty and comes before the last non-empty urn. To distinguish clearly from the maximal gaps mentioned before, we could call this "gaps in the weak sense"; of course, for the terminology "gap-free" both versions amount asymptotically to the same.*
>
> *In this paper, we study these* gaps in the weak sense. *However, since the name "gaps" should be reserved for the version studied by P. Hitczenko and A. Knopfmacher [11], we will use terminology such as "the number of unoccupied urns," or similar ones.*

Hitczenko and Knopfmacher analyze the quantity

$$p_n(0) := \mathbb{P}[\text{All urns are occupied up to the maximal non-empty urn}].$$

Recently, Goh and Hitczenko [9] have continued the study of gaps in the "maximal" sense, as described before.

In our paper, we analyze the probability $p_n(r)$ of having r empty urns, the moments of the total number of empty urns and some other parameters.

As a link to more practically oriented research, we mention *probabilistic counting* [7], which can be seen in the context of our empty urn discussion.

The case $p = 1/2$ has a particular interest: it is related to the compositions of integers, see [12].

It is intuitively not at all clear, but nevertheless true, that the quantities that we analyze for general p, simplify for the special choice of $p = 1/2$. This produces identities, since a complicated expression simplifies for a special choice of the parameter. Now, one gets these identities "for free," since two different approaches must eventually lead to the same result. Nevertheless, we believe that there is a genuine interest in producing independent (analytic) proofs for these simplifications. To give a flavour of such identities,

$$\int_0^\infty e^{-y}\left[\sum_{j=0}^{\infty}(-1)^{\nu(j)}e^{-jy}\right]\frac{dy}{y} = \frac{\ln 2}{2},$$

[1] Here we use the indicator function ('Iverson's notation') proposed by Graham et al. [10].

where $\nu(j)$ is the number of ones in the binary representation of j.

For those readers who know the literature on analysis of algorithms well, we provide the following *informal* comment that might be helpful and might be compared with the one described in [20, 21]: Since a variance cannot be negative and the main term fluctuates around zero, the fluctuation must be identical to zero and the Fourier coefficients must be equal to zero. Now, such a combinatorial argument is nice and sweet when it applies! But there are situations as well, when one has to *compute* the Fourier coefficients, and that is at the same level of complexity as to prove in the other instances that they are zero.

The probabilities and moments we compute will have an asymptotic expansion of the form

$$D \sim D^* + D_P(\log n),$$

where the dominant term D^* is independent of n and D_P is a periodic function, with 0 mean, of period 1 and usually of small amplitude, of order 10^{-5}.

We also provide some asymptotics for p going to 0 and 1.

We will obtain several limiting distributions. To show that the limiting moments are equivalent to the moments of the limiting distributions, we need a suitable rate of convergence. This is related to a uniform integrability condition (see Loève [15, Section 11.4]). For the kind of limiting distributions we consider here, the rate of convergence is analyzed in detail in [17] and [19]. As a byproduct, we also derive some interesting combinatorial identities.

Here is the plan of the paper: Section 2 deals with the probability that a sequence of length n has a fixed number r of empty urns; this probability is asymptotically given by some function $p^*(r)$ plus some small oscillations. When $p = q = 1/2$, everything simplifies drastically; there are no oscillations. We also study what happens when $p \to 0$ (or, equivalently, when $q \to 1$) and when $p \to 1$.

Section 3 deals with the random variable X: 'number of empty urns.' The moments are evaluated asymptotically. There is also a technical study of the distribution of X, when $p \to 0$.

Section 4 considers the samples that have r empty urns. Among those, the random variable 'largest non-empty urn' is considered. Again, we are able to evaluate all moments of its distribution asymptotically.

In a final Section 5 we briefly sketch the concept we nickname 'super-gap-free.' A sample is super-gap-free, if it is gap-free and each non-empty urn contains at least 2 items (and d-super-gap-free, if they contain $\geq d$ items). For the instance $p = q = 1/2$, we sketch how the asymptotic probability (apart from small oscillations) that a sample is d-super-gap-free can be computed. We leave further studies (general parameter p, higher moments, fixed number of empty urns, etc.) to the interested reader.

Note that not all details of our computations are given in our paper. They can be found in the long version of our work [18].

The procedure to obtain our results in this paper is as follows: Some analytic expressions are derived, from which is becomes clear how to derive all moments in a more or less automatic way. Then we derive the special identities that are related to the case $p = 1/2$, and then we analyze what happens when the parameter p tends to 0 or 1.

Let us finally note that most of the literature on urns concerns the uniform occupancy distribution, while our work focuses on distribution of balls in urn j given by pq^{j-1}. Let us mention two recent papers dealing with the non-uniform case, where many references can be found: Bogachev et al. [5] and Hwang and Janson [13].

2 The probability of empty urns

In this section, we consider the case of r empty urns, when $n \to \infty$.

2.1 No empty urns

In this section, we analyze $p_n(0)$. We first prove asymptotic independence of urns, by a Poissonization-depoissonization technique. Then we obtain a new asymptotic form for $p_n(0)$ and give asymptotic expressions when $p \to 1$ and $p \to 0$. Finally, we consider the particular case $p = 1/2$.

2.1.1 The general case

Let us recall that $p_n(r) := \mathbb{P}$[There are r empty urns up to the maximal non-empty urn].
So $p_n(0) = \mathbb{P}$[All urns are occupied up to the maximal non-empty urn]. Assume in the sequel that this maximal non-empty urn is urn k. We will use the following notations:

$$\begin{aligned}
Q &:= 1/q, \\
L &:= \ln 1/q = \ln Q, \\
n^* &:= np/q, \\
\log &:= \log_Q, \\
\alpha &:= q/p, \\
X &:= \text{the total number of empty urns.}
\end{aligned}$$

In Hitczenko and Louchard [12], the case $p = 1/2$ was analyzed. It was proved that asymptotically, the urns become independent. To prove asymptotic independence in our case, we consider the generating function $F_n(z)$ of the $p_n(r)$'s: $F_n(z) := \mathbb{E}(z^X)$.

Theorem 2.1

$$F_n(z) \sim \sum_{k=1}^{\infty} \left[1 - e^{-n^*/Q^k}\right] e^{-\alpha n^*/Q^k} \prod_{w=1}^{k-1} \left[(z-1)e^{-n^* Q^w/Q^k} + 1\right] + \mathcal{O}(n^{\beta_1 - 1}), \quad n \to \infty,$$

uniformly for $z \in S$, $S := \{|z| < 1, |z-1| \le L_1\}$, $q \le e^{-L_2}$, *with* $\beta_1 = L_1/L_2 < 1$, $0 < L_1 < L_2$.

Proof
We will use the Poissonization method (see, for instance Jacquet and Szpankowski [14] for a general survey). Assume that the last full urn is urn k. First we must consider the empty urns before urn k. Next, all urns after urn k must be empty. If we Poissonize the number

of balls (i.e., the number of R.V. here), with parameter τ, we have independence of cells occupation. We obtain

$$F(z,\tau^*) := e^{-\tau}\sum_n \frac{\tau^n}{n!}F_n(z) = \sum_{k=1}^{\infty}\left[1-e^{-\tau^*/Q^k}\right]e^{-\alpha\tau^*/Q^k}\prod_{w=1}^{k-1}\left[(z-1)e^{-\tau^* Q^w/Q^k}+1\right],$$

with $\tau^* := \tau p/q$. Hence, by Cauchy's integral theorem, we obtain

$$F_n(z) = \frac{n!}{2\pi\mathbf{i}}\int_\Gamma F(z,\tau^*)e^\tau d\tau/\tau^{n+1},$$

where Γ is inside the analyticity domain of $F(z,\tau^*)$ and encircles the origin.

We will use [23, Thm.10.3 and Cor.10.17]. Assume that in a linear cone $S_\theta = \{z : |\arg(z)| \le \theta, 0<\theta<\pi/2\}$, the following two conditions simultaneously hold for all z in a set S:

(I): For $\tau \in S_\theta$ and some reals B, $R>0$, β_1,

$$|\tau| > R \implies |F(z,\tau^*)| \le B|\tau|^{\beta_1}.$$

(O) For $\tau \notin S_\theta$ and $A, \beta_2 < 1$,

$$|\tau| > R \implies |F(z,\tau^*)e^\tau| \le A\exp(\beta_2\tau).$$

Then

$$F_n(z) = F(z,n^*) + \mathcal{O}(n^{\beta_1-1})$$

uniformly for $z \in S$.

Let us first check (O). Let $|z|<1$. First of all

$$|e^\tau F(z,\tau^*)| \le \sum\frac{|\tau|^n}{n!} = e^{|\tau|}.$$

Moreover, we have

$$\begin{aligned}
F(z,\tau^*) &= \left[1-e^{-\tau^*/Q}\right]e^{-\alpha\tau^*/Q} + \sum_{k=2}^{\infty}\left[1-e^{-\tau^*/Q^k}\right]e^{-\alpha\tau^*/Q^k}\prod_{w=1}^{k-1}\left[(z-1)e^{-\tau^*Q^w/Q^k}+1\right] \\
&= \left[1-e^{-\tau^*/Q}\right]e^{-\alpha\tau^*/Q} \\
&+ \sum_{k=1}^{\infty}\left[1-e^{-\tau^*/Q^{k+1}}\right]e^{-\alpha\tau^*/Q^{k+1}}\prod_{w=1}^{k-1}\left[(z-1)e^{-\tau^*Q^w/Q^{k+1}}+1\right]\left[(z-1)e^{-\tau^*/Q}+1\right] \\
&= \left[1-e^{-\tau^*/Q}\right]e^{-\alpha\tau^*/Q} + F(z,\tau^*/Q)\left[(z-1)e^{-\tau^*/Q}+1\right].
\end{aligned}$$

This gives

$$\begin{aligned}
e^\tau F(z,\tau^*) &= e^{p\tau}-1+e^\tau F(z,\tau^* q)\left[(z-1)e^{-\tau p}+1\right] \\
&= e^{p\tau}-1+e^{\tau q}F(z,\tau^* q)\left[(z-1)+e^{\tau p}\right].
\end{aligned}$$

But

$$|e^{\tau p}| = e^{\Re(\tau)p}, \quad |e^{\tau q}| = e^{\Re(\tau)q}, \quad \Re(\tau) \le |\tau|\cos(\theta),$$

so

$$|e^\tau F(z,\tau^*)| \le 1 + e^{|\tau| p\cos(\theta)} + e^{|\tau| q}\left[1+e^{|\tau|p\cos(\theta)}\right] \le 2e^{|\tau|(q+p\cos(\theta))},$$

and (O) is satified with $\beta_2 = q + p\cos(\theta) < 1$.

Now we check condition (I). Set $\tau = x + \mathbf{i}y$, $x>0$, $\tau^* = x^* + \mathbf{i}y^*$. We consider the sum in $F(z,\tau^*)$ that we split into two parts.

- $k \le \lfloor \log x \rfloor$. We have, with $|z-1|\le\rho$,

  $$\left|\prod_{w=1}^{k-1}\left[(z-1)e^{-\tau^* Q^w/Q^k}+1\right]\right| = \left|\prod_{u=1}^{k-1}\left[(z-1)e^{-\tau^* q^u}+1\right]\right| \le \prod_{u=1}^{k-1}|1+\rho| \le e^{\rho k}.$$

  Also

  $$|1-e^{-\tau^*/Q^k}| \le 2,$$

  and

  $$|e^{-\alpha\tau^*/Q^k}| = |e^{-\tau q^k}| = e^{-xq^k}.$$

  But

  $$S_1 := 2\sum_{1}^{\lfloor \log x\rfloor} e^{-xq^k}e^{\rho k} = 2\sum_{1}^{\lfloor \log x\rfloor} e^{-xQ^{-k}}e^{\rho k}.$$

  Set $\eta := L(k-\log x)$. For $|\tau|$ large, with $\gamma_1 = \rho/L$,

  $$S_1 \sim 2\int_{-\infty}^{0}\exp(-e^{-\eta})x^{\gamma_1}e^{\gamma_1\eta}d\eta/L \le 2|\tau|^{\gamma_1}\int_1^\infty e^{-y}\frac{dy}{y^{1+\gamma_1}} = C_1|\tau|^{\gamma_1}.$$

- $\lfloor \log x\rfloor \le k < \infty$. We have

  $$|e^{-\alpha\tau^*/Q^k}| \le 1,$$

  and, by standard algebra,

  $$\begin{aligned}|1-e^{-\tau^*/Q^k}| &\le 1-e^{-x^*/Q^k} + \frac{1-\cos(y^*/Q^k)}{\exp(x^*/Q^k)-1} \le 1-e^{-x^*/Q^k} + \frac{y^{*2}/Q^{2k}}{2(\exp(x^*/Q^k)-1)}\\ &\le 1-e^{-x^*/Q^k} + \frac{\tan^2(\theta)}{2}\frac{x^{*2}/Q^{2k}}{\exp(x^*/Q^k)-1}.\end{aligned}$$

  Consider

  $$S_2 := \sum_{\lfloor \log x^*\rfloor}^{\infty}\left[1-e^{-x^*/Q^k}\right]e^{\rho k} + \frac{\tan^2(\theta)}{2}\sum_{\lfloor \log x^*\rfloor}^{\infty}\frac{x^{*2}/Q^{2k}}{\exp(x^*/Q^k)-1}e^{\rho k}.$$

  Set $\eta := L(k-\log x^*)$ (the lower index in the sum, compared with $\lfloor \log x\rfloor$, only introduces an extra constant contribution). We derive, with $\gamma_2 = \rho/L$,

  $$S_2 \sim x^{*\gamma_2}\int_0^\infty[1-\exp(-e^{-\eta})]e^{\gamma_2\eta}d\eta/L + x^{*\gamma_2}\frac{\tan^2(\theta)}{2}\int_0^\infty \frac{e^{-2\eta}}{\exp(e^{-\eta})-1}e^{\gamma_2\eta}d\eta/L \le C_2|\tau|^{\gamma_2},$$

  if $\rho < L$.

So, finally

$$S_1 + S_2 \le C_1|\tau|^{\gamma_1} + C_2|\tau|^{\gamma_2},$$

and condition (I) is satisfied with $\beta_1 = \rho/L$. Note that $\rho < L$ is equivalent to $e^\rho < Q$, so this is automatically satified if $q < e^{-2} = 0.1353\ldots$. If $q \ge e^{-2}$, we must choose $\rho < L$.

The theorem is now immediately derived. ∎

Let us consider (the product is now extended to ∞)

$$G_n(z) := \sum_{k=1}^{\infty} \left[1 - e^{-n^*/Q^k}\right] e^{-\alpha n^*/Q^k} \prod_{w=1}^{\infty} \left[(z-1)e^{-n^* Q^w/Q^k} + 1\right].$$

We want to replace $F_n(z)$ as given by Theorem 2.1 by $G_n(z)$. We see that this amounts to adding, to X, an independent RV $\tilde{X}$ such that $\tilde{X}$ is the sum of independent Bernoulli RV ξ_i, $i = 0, 1, 2, \ldots$, with

$$\mathbb{P}[\xi_i = 1] = e^{-n^* Q^i}.$$

We have

$$\mathbb{E}(\tilde{X}) = \sum_{i=0}^{\infty} e^{-n^* Q^i} \sim e^{-n^*}, \quad n \to \infty.$$

Similarly, the variance is given by

$$\mathbb{V}(\tilde{X}) = \sum_{i=0}^{\infty} e^{-n^* Q^i} \left(1 - e^{-n^* Q^i}\right) \sim e^{-n^*}, \quad n \to \infty.$$

By Chebyshev's inequality,

$$\mathbb{P}\left[\tilde{X} - e^{-n^*} \ge \beta e^{-n^*/2}\right] \le \frac{1}{\beta^2},$$

which shows that $\tilde{X}$ is exponentially small. For instance, if we set $\beta e^{-n^*/2} + e^{-n^*} = 1$, this leads to $\beta \sim e^{n^*/2}$ and $\mathbb{P}(\tilde{X} \ge 1) \le e^{-n^*}$. So, from now on, we will always use $G_n(z)$.

The function $G_n(z)$ is a harmonic sum, so we define

$$\Lambda(z,s) := \int_0^{\infty} y^{s-1} e^{-\alpha y} \prod_{w=1}^{\infty} \left[1 + (z-1)e^{-yQ^w}\right](1 - e^{-y})dy, \tag{1}$$

and the Mellin transform of the sum is (for a good reference on Mellin transforms, see Flajolet et al. [6] or Szpankowski [23])

$$\frac{Q^s}{1 - Q^s}\Lambda(z,s). \tag{2}$$

The fundamental strip of (2) is $s \in \langle -1, 0 \rangle$. We see that, for $\Re(s) > -1$, all poles of $\frac{Q^s}{1-Q^s}\Lambda(z,s)$ are simple poles. Indeed, for $|z| < 1$,

$$\left|\prod_{w=1}^{\infty} \left[1 + (z-1)e^{-yQ^w}\right]\right| < 1, \tag{3}$$

and

$$\int_0^\infty e^{-\alpha y} y^{s-1} \left[1 - e^{-y}\right] dy = \Gamma(s) \left[\frac{1}{\alpha^s} - \frac{1}{(\alpha+1)^s}\right],$$

which has no poles for $\Re(s) > -1$.

The poles of $\frac{Q^s}{1-Q^s}\Lambda(z,s)$ are thus given by $s = 0$, $s = \chi_l$, with $\chi_l := 2l\pi\mathbf{i}/L$, $l \in \mathbb{Z} \setminus \{0\}$.

Also, we have a "slow increase property": by (3), we have

$$|\Lambda(z,s)| \le \left| \int_0^\infty e^{-\alpha y} y^{s-1} \left[1 - e^{-y}\right] dy \right| = \left| \Gamma(s) \left[\frac{1}{\alpha^s} - \frac{1}{(\alpha+1)^s}\right] \right|$$

and $\Gamma(s)$ decreases exponentially towards $\mathbf{i}\infty$. This is necessary to allow moving the line of integration to the right (see details in [23]).

Using

$$G_n(z) = \frac{1}{2\pi\mathbf{i}} \int_{C_2 - \mathbf{i}\infty}^{C_2 + \mathbf{i}\infty} \frac{Q^s}{1-Q^s} \Lambda(z,s)(n^*)^{-s} ds, \quad -1 < C_2 < 0,$$

the asymptotic expression of $G_n(z)$ (for large n) is obtained by moving the line of integration to the right, for instance to the line $\Re(s) = C_4 > 0$, taking residues into account (with a negative sign). This gives

$$G_n(z) = -\mathrm{Res}\Big[\frac{Q^s}{1-Q^s}\Lambda(z,s)(n^*)^{-s}\Big]\Big|_{s=0} - \sum_{l\neq 0} \mathrm{Res}\Big[\frac{Q^s}{1-Q^s}\Lambda(z,s)(n^*)^{-s}\Big]\Big|_{s=\chi_l} + \mathcal{O}\big(n^{-C_4}\big),$$

for $n \to \infty$, uniformly on $|z| \le 1 - \rho$, $\rho > 0$.

This leads to

$$G_n(z) \sim G^*(z) + \frac{1}{L} \sum_{l \neq 0} \Lambda(z, \chi_l) e^{-2l\pi\mathbf{i}\log n^*}, \quad n \to \infty, \tag{4}$$

with

$$G^*(z) := \int_0^\infty e^{-\alpha y} \prod_{w=1}^\infty \left[1 + (z-1)e^{-yQ^w}\right] (1 - e^{-y}) \frac{dy}{Ly}. \tag{5}$$

Note that $G^*(z)$ is independent of n^*.

2.1.2 Analysis of $p_n(0)$

To analyze $p_n(0)$, we can proceed as in Louchard and Prodinger [17, Section 4.8 and 5.9] and in Louchard, Prodinger and Ward [19], where it is shown that we can replace Binomials by Poisson distributions and where the rate of convergence is analyzed in detail. This entails that the limiting moments are given by the moments of the limiting distribution (this is related to a uniform integrability condition, see Loève [15, Sec.11.4]).

For instance,

$$\mathbb{P}(X_i = 0) = (1 - pq^{i-1})^n \sim e^{-npq^{i-1}} = e^{-n^* q^i}, \quad \text{for } i = \log n + \mathcal{O}(1).$$

Note that Theorem 2.1 already gives a Poissonized form (with asymptotic independence) of the limiting distribution. See Barbour et al. [4] for more on Poisson approximations.

Assume again that the last full urn is urn k. Here we must have all urns full before and including urn k. Next, all urns after urn k must be empty. This leads, asymptotically, to

$$\begin{aligned} p_n(0) &\sim \sum_{k=1}^{\infty} \prod_{i=0}^{\infty} \left[1 - e^{n^* Q^i / Q^k}\right] \prod_{w=1}^{\infty} \left[e^{-n^*/Q^{w+k}}\right] \\ &= \sum_{k=1}^{\infty} \prod_{i=0}^{\infty} \left[1 - e^{-n^* Q^i / Q^k}\right] e^{-\alpha n^*/Q^k}, \quad n \to \infty \end{aligned} \tag{6}$$

with $\alpha = q/p$.

This is again a harmonic sum, so we define

$$\varphi(0, s) := \int_0^{\infty} y^{s-1} e^{-\alpha y} \prod_{i=0}^{\infty} \left[1 - e^{-y e^{Li}}\right] dy, \tag{7}$$

and proceeding as previously, we obtain

$$p_n(0) = -\mathrm{Res}\Big[\frac{Q^s}{1 - Q^s} \varphi(0, s)(n^*)^{-s}\Big]\Big|_{s=0} - \sum_{l \neq 0} \mathrm{Res}\Big[\frac{Q^s}{1 - Q^s} \varphi(0, s)(n^*)^{-s}\Big]\Big|_{s=\chi_l} + \mathcal{O}\big(n^{-C_4}\big), \quad n \to \infty.$$

This leads to

$$\begin{aligned} p_n(0) \quad &\sim \quad p^*(0) + \frac{1}{L} \sum_{l \neq 0} \varphi(0, \chi_l) e^{-2l\pi \mathbf{i} \log n^*}, \quad n \to \infty, \text{ with} \end{aligned} \tag{8}$$

$$p^*(0) \quad = \quad \int_0^{\infty} e^{-\alpha y} \prod_{i=0}^{\infty} \left[1 - e^{-y e^{Li}}\right] \frac{dy}{Ly}. \tag{9}$$

The expression (8), with (9), is direct and simpler than the one given in Hitczenko and Knopfmacher [11], which depends on the previous values $p_j(0)$.

One way to compute $p^*(0)$ numerically is the following: we have

$$\int_0^{\infty} e^{-\alpha y} \left[1 - e^{-y}\right] \frac{dy}{y} = \ln(\alpha + 1) - \ln(\alpha) = \ln\Big(\frac{1 + \alpha}{\alpha}\Big) = \ln\Big(1 + \frac{1}{\alpha}\Big),$$

so, if we set

$$\begin{aligned} f_0(\alpha) &= \ln(\alpha + 1) - \ln(\alpha), \\ f_1(\alpha) &= f_0(\alpha) - f_0(\alpha + Q), \\ f_i(\alpha) &= f_{i-1}(\alpha) - f_{i-1}(\alpha + Q^i), \end{aligned} \tag{10}$$

a numerical asymptotic expression for $p^*(0)$ is given by $f_i(\alpha)/L$. Indeed, the uniform convergence of the product in (9) entails the convergence of $f_i(\alpha)$. An experiment with Maple works quite well for $p \geq .5$, with $i = 12$, but for $p < 0.5$, i must be larger and larger and we have inaccuracies due to loss of significant digits. Let us try to simplify the iteration. For any integer j, denote by $a[.]$ the vector of its binary representation. We set $\nu(j) = \sum_{u=0}^{\infty} a[u]$

and $g(j) := \sum_{u=0}^{\infty} a[u]Q^{u+1}$; $\nu(j)$ denotes, as usual, the number of ones in the binary representation of j. It is easily checked that

$$f_i(\alpha) = \sum_{j=0}^{2^i-1} (-1)^{\nu(j)} \ln\left(\frac{\alpha+1+g(j)}{\alpha+g(j)}\right). \tag{11}$$

This improves the computation speed for $p < 0.5$, but the main interest is to give the following:

Asymptotic expressions for $p^*(0)$.

1. Let $p = 1 - \varepsilon$.

   Expanding $f_i(\alpha)$, we obtain the following asymptotic expansion

   $$p^*(0) \sim 1 + \varepsilon/\ln(\varepsilon) - \varepsilon^2/(2\ln(\varepsilon)) + \varepsilon^3/(3\ln(\varepsilon)) + \cdots, \quad \varepsilon \to 0.$$

   For $\varepsilon = 0.01$, we obtain, by numerical integration of (8), $p^*(0) = 0.9978393126\ldots$ and the asymptotic expansion above gives an error of the order 10^{-10}.

2. If we let $p = \varepsilon$, we first bound

   $$\int_0^A e^{-\alpha y} \prod_{i=0}^{\infty} \left[1 - e^{-ye^{Li}}\right] \frac{dy}{Ly},$$

   and

   $$\int_B^{\infty} e^{-\alpha y} \prod_{i=0}^{\infty} \left[1 - e^{-ye^{Li}}\right] \frac{dy}{Ly},$$

   for some constants A, B that will be fixed later on. We have, for $l^* = 1/\varepsilon$,

   $$\begin{aligned}
   \int_0^A e^{-\alpha y} \prod_{i=0}^{\infty} \left[1 - e^{-ye^{Li}}\right] \frac{dy}{Ly} &\sim \int_0^A e^{-y/\varepsilon}(1-e^{-y}) \prod_{l=1}^{\infty} \left[1 - e^{-ye^{\varepsilon l}}\right] \frac{dy}{Ly} \\
   &\le \int_0^A e^{-y/\varepsilon}(1-e^{-y}) \prod_{l=1}^{l^*} \left[1 - e^{-ye^{\varepsilon l}}\right] \frac{dy}{Ly} \\
   &\le \int_0^A (1-e^{-y})(1-e^{-Be})^{1/\varepsilon} \frac{dy}{Ly} = \mathcal{O}\left(e^{-2/\varepsilon}/\varepsilon\right)
   \end{aligned}$$

   if we choose $A = -\ln(1-e^{-2})/e = 0.053\ldots$.

   Similarly

   $$\begin{aligned}
   \int_B^{\infty} e^{-\alpha y} \prod_{i=0}^{\infty} \left[1 - e^{-ye^{Li}}\right] \frac{dy}{Ly} &\le \int_B^{\infty} e^{-y/\varepsilon} \frac{dy}{Ly} \\
   &\sim e^{-B/\varepsilon} \frac{\varepsilon}{BL} = \mathcal{O}\left(e^{-2/\varepsilon}\right)
   \end{aligned}$$

   if we choose $B = 2$.

Now, with the Euler-Maclaurin summation formula,

$$\begin{aligned}\ln\left(\prod_{i=0}^{\infty}\left[1-e^{-ye^{Li}}\right]\right) &\sim \sum_0^{\infty}\ln\left(1-e^{-ye^{\varepsilon l}}\right)\\ &\sim \int_0^{\infty}\ln\left(1-e^{-ye^{\varepsilon l}}\right)dl+\frac{1}{2}\ln(1-e^{-y})+\mathcal{O}(\varepsilon),\end{aligned}$$

uniformly on $y\in(A,B)$.

Neglecting, in first approximation, the last correction, we derive

$$\begin{aligned}p^*(0) &\sim \int_A^B \exp\left[-y/\varepsilon+\int_0^{\infty}\ln\left(1-e^{-ye^{\varepsilon l}}\right)dl\right]\frac{dy}{yL}\\ &= \int_A^B \exp\left[-y/\varepsilon+I_0(y)/\varepsilon\right]\frac{dy}{yL},\quad \text{with}\\ I_0(y) &:= \int_1^{\infty}\ln\left(1-e^{-yv}\right)\frac{dv}{v}.\end{aligned}$$

So we can apply the saddle point method, see [8]. The saddle point $\tilde{y}$ is given by $I_0'(\tilde{y})=1$, i.e., $\tilde{y}=\ln(2)$. We derive

$$\begin{aligned}I_0(\tilde{y}) &= -0.4592756884\ldots,\\ I_2(y) &:= I_0''(y)=\left[e^y\ln(e^y-1)-e^y y-\ln(e^y-1)\right]/\left[y^2(e^y-1)\right],\\ I_2(\tilde{y}) &= -2/\ln(2)=-2.885390082\ldots\end{aligned}$$

Now, we obtain

$$\begin{aligned}p^*(0) &\sim e^{[-\tilde{y}+I_0(\tilde{y})]/\varepsilon}\int_0^{\infty}e^{(y-\tilde{y})^2I_2(\tilde{y})/(2\varepsilon)}\frac{dy}{yL}\\ &\sim e^{[-\tilde{y}+I_0(\tilde{y})]/\varepsilon}\frac{\sqrt{2\pi}}{\tilde{y}\sqrt{\varepsilon|I_2(\tilde{y})|}}=\exp[-1.152422869\ldots/\varepsilon]\frac{2.128934039\ldots}{\sqrt{\varepsilon}}. \quad (12)\end{aligned}$$

Indeed, first, $|-\tilde{y}+I_0(\tilde{y})|<2$, and next,

$$\int_B^{\infty}e^{(y-\tilde{y})^2I_2(\tilde{y})/(2\varepsilon)}\frac{dy}{yL}\sim e^{-C/\varepsilon}$$

by Laplace's method, with $C=-(B-\tilde{y})^2I_2(\tilde{y})/2$ (and similarly for A).

For $\varepsilon=0.05$, we obtain, by numerical integration of (8), $p^*(0)=0.1646576705\ldots 10^{-8}$ and the asymptotic expansion above gives $0.9\ldots 10^{-9}$. (Of course, this is only a first order expression, which could be refined, for instance using the neglected corrections.)

Let us now return to $\varphi(0,s)$ as given by (7). We want to obtain an asymptotic expression for this quantity. Starting from

$$\int_0^{\infty}e^{-\alpha y}y^{s-1}\left[1-e^{-y}\right]dy=\Gamma(s)\left[\frac{1}{\alpha^s}-\frac{1}{(\alpha+1)^s}\right],$$

we set

$$\begin{aligned}
\phi_0(\alpha, s) &= \Gamma(s)\left[\frac{1}{\alpha^s} - \frac{1}{(\alpha+1)^s}\right], \\
\phi_i(\alpha, s) &= \phi_{i-1}(\alpha, s) - \phi_{i-1}(\alpha + Q^i, s),
\end{aligned}$$

and an asymptotic expression for $\varphi(0, s)$ is given by $\phi_i(\alpha, s)$, i large. Again, the uniform convergence of the product in (7) entails the convergence of $\phi_i(\alpha, s)$. Note that, when $s \to 0$, we recover the previous expression (10).

2.1.3 Simplifications for $p = 1/2$

This case has a particular interest: it is related to the compositions of integers, see [12]. We have the simplification

$$\varphi(0, s) := \int_0^\infty y^{s-1} e^{-y}\left[\sum_{j=0}^{\infty}(-1)^{\nu(j)} e^{-jy}\right] dy = \Gamma(s)\sum_{j=0}^{\infty}(-1)^{\nu(j)}\frac{1}{(j+1)^s}, \tag{13}$$

where $\nu(j)$ is the number of ones in the binary representation of j. In Hitczenko and Knopfmacher [11] it is shown that $p_n(0) = 1/2$, for all n. Of course we must have

$$p^*(0) = \int_0^\infty e^{-y}\left[\sum_{j=0}^{\infty}(-1)^{\nu(j)} e^{-jy}\right]\frac{dy}{Ly} = \frac{1}{2}. \tag{14}$$

This is not only important as a check of our asymptotic expressions, but also it will lead to some interesting identities, and to simple proofs of some constant values found in the literature.

The function

$$M(s) = \sum_{j\ge 0}(-1)^{\nu(j)}\frac{1}{(j+1)^s} \tag{15}$$

can be treated similarly to the classical entire function $N(s)$ which is the analytic continuation of

$$\sum_{j\ge 1}(-1)^{\nu(j)}/j^s;$$

see, for instance, Flajolet and Martin [7], Louchard and Prodinger [17]: We group 2 terms together (resp. 4), for analytic continuation. Now, we have for $s \to 0$,

$$\frac{1}{(2j+1)^s} - \frac{1}{(2j+2)^s} \sim s\ln\frac{2j+2}{2j+1}, \qquad \Gamma(s) \sim \frac{1}{s},$$

so

$$\Gamma(s)M(s) \sim \sum_{j\ge 0}(-1)^{\nu(j)}\ln\frac{2j+2}{2j+1} = \ln\prod_{j\ge 0}\left(\frac{2j+2}{2j+1}\right)^{(-1)^{\nu(j)}}.$$

However, we find in Allouche and Shallit [3] (compare also [22, 1, 2]) that

$$\prod_{j\ge 0}\left(\frac{2j+1}{2j+2}\right)^{(-1)^{\nu(j)}} = \frac{1}{\sqrt{2}}. \tag{16}$$

We have eventually

$$\ln(\sqrt{2}\,)/L = \log(\sqrt{2}\,) = \frac{1}{2},$$

proving (14). Note that, if we let $p = 1/2$ in (11), we obtain another simple proof of (16).

Now we will also prove directly that the periodic component is null for $p = 1/2$. This leads to an interesting identity:

$$M(\chi_l) = \sum_{j=0}^{\infty}(-1)^{\nu(j)}\frac{1}{(j+1)^{\chi_l}} = 0. \tag{17}$$

We introduce a technique that will be frequently used in the sequel. Let

$$f(x) = \sum_{j\geq 0}(-1)^{\nu(j)}\frac{1}{(j+x)^{\chi}},$$

where $2^{\chi} = 1$. Then,

$$\begin{aligned} f(x) &= \sum_{j\geq 0}(-1)^{\nu(2j)}\frac{1}{(2j+x)^{\chi}} + \sum_{j\geq 0}(-1)^{\nu(2j+1)}\frac{1}{(2j+x+1)^{\chi}} \\ &= \sum_{j\geq 0}(-1)^{\nu(j)}\frac{1}{(2j+x)^{\chi}} - \sum_{j\geq 0}(-1)^{\nu(j)}\frac{1}{(2j+x+1)^{\chi}} \\ &= \sum_{j\geq 0}(-1)^{\nu(j)}\frac{1}{(j+\frac{x}{2})^{\chi}} - \sum_{j\geq 0}(-1)^{\nu(j)}\frac{1}{(j+\frac{x+1}{2})^{\chi}} \\ &= f(\tfrac{x}{2}) - f(\tfrac{x+1}{2}). \end{aligned}$$

We cannot strictly plug in $x = 0$ as it stands, but we subtract one term $(1/x^{\chi})$ from each side:

$$\sum_{j\geq 1}(-1)^{\nu(j)}\frac{1}{(j+x)^{\chi}} = \sum_{j\geq 1}(-1)^{\nu(j)}\frac{1}{(j+\frac{x}{2})^{\chi}} - \sum_{j\geq 0}(-1)^{\nu(j)}\frac{1}{(j+\frac{x+1}{2})^{\chi}},$$

and now it is possible to plug in $x = 0$. Note that $\chi \neq 0$. Furthermore, using the technique of "grouping 2 (or 4, or 8, ...) terms together" as used by Flajolet and Martin [7] and also Section 4, removes all convergence issues and produces an analytic continuation of $f(x)$.

So, we look at

$$f(x) - \frac{1}{x^{\chi}} = f(\tfrac{x}{2}) - \frac{1}{x^{\chi}} - f(\tfrac{x+1}{2}).$$

Now we can plug in $x = 0$, which amounts to take the limit $x \to 0$ in

$$f(x) = f(\tfrac{x}{2}) - f(\tfrac{x+1}{2}),$$

and we find that $f(\frac{1}{2}) = 0$. Now plugging in $x = 1$, we find

$$f(1) = f(\tfrac{1}{2}) - f(1),$$

and so $f(1) = 0$, as desired.

Of course,

$$\lim_{y\to 0}\sum_{j=0}^{\infty}(-1)^{\nu(j)}e^{-jy} = 0, \quad \text{as } N(0) = -1.$$

This absence of fluctuations is not uncommon: it is also the case for the variance of the number of distinct part sizes in composition of large integers [12], the mean of adaptative sampling [16], the variance of Patricia Tries [20, 21].

To summarize, we have:

Theorem 2.2 *The probability $p_n(0)$ that there is no empty urn, is asymptotically given by*

$$p_n(0) \sim p^*(0) + \frac{1}{L}\sum_{l\neq 0}\varphi(0,\chi_l)e^{-2l\pi\mathbf{i}\log n^*}, \quad n\to\infty, \qquad \textit{with}$$
$$p^*(0) = \int_0^\infty e^{-\alpha y}\prod_{i=0}^{\infty}\left[1-e^{-ye^{Li}}\right]\frac{dy}{Ly}.$$

The periodic component is given via

$$\varphi(0,s) := \int_0^\infty y^{s-1}e^{-\alpha y}\prod_{i=0}^{\infty}\left[1-e^{-ye^{Li}}\right]dy.$$

In the symmetric case $p = q = 1/2$, these expressions simplify, and $p_n(0) = 1/2$.

2.2 More empty urns

In this section, we analyze $p_n(r)$, again including asymptotic expressions when $p \to 1$ and $p \to 0$ and the particular case $p = 1/2$. An inclusion-exclusion argument leads us to some curious identities.

2.2.1 The general case

We have $p_n(1) := \mathbb{P}$[There is one empty urn up to the maximal non-empty urn]. This leads, asymptotically, to

$$p_n(1) \sim \sum_{k=1}^{\infty}\prod_{i=0}^{\infty}\left[1-e^{-n^*Q^i/Q^k}\right]e^{-\alpha n^*/Q^k}\sum_{u=1}^{\infty}\frac{e^{-n^*Q^u/Q^k}}{1-e^{-n^*Q^u/Q^k}}, \quad n\to\infty.$$

Indeed, assume as usual that the last full urn is urn k. Here we must have all urns $k-i$ full before k except for one urn, $k-u$ say. This amounts to replacing the term $1-e^{-n^*Q^u/Q^k}$ by $e^{-n^*Q^u/Q^k}$. Next, all urns after urn k must be empty. Set

$$\varphi(1,s) := \int_0^\infty y^{s-1}e^{-\alpha y}\prod_{i=0}^{\infty}\left[1-e^{-ye^{Li}}\right]\sum_{u=1}^{\infty}\frac{e^{-ye^{Lu}}}{1-e^{-ye^{Lu}}}dy. \tag{18}$$

This leads to

$$p_n(1) \sim p^*(1) + \frac{1}{L}\sum_{l\neq 0}\varphi(1,\chi_l)e^{-2l\pi\mathbf{i}\log n^*}, \quad n\to\infty,$$

with

$$p^*(1) = \int_0^\infty e^{-\alpha y}\prod_{i=0}^{\infty}\left[1-e^{-ye^{Li}}\right]\sum_{u=1}^{\infty}\frac{e^{-ye^{Lu}}}{1-e^{-ye^{Lu}}}\frac{dy}{Ly}. \tag{19}$$

The case of 2 or more empty urns is easily generalized, by introducing more sums. For instance, $p^*(2)$ is given by

$$p^*(2) = \int_0^\infty e^{-\alpha y} \prod_{i=0}^\infty \left[1 - e^{-ye^{Li}}\right] \sum_{1 \le u_1 < u_2} \frac{e^{-ye^{Lu_1}}}{1 - e^{-ye^{Lu_1}}} \frac{e^{-ye^{Lu_2}}}{1 - e^{-ye^{Lu_2}}} \frac{dy}{Ly}.$$

We have the following result

Theorem 2.3 *The probability $p_n(r)$ that there are r empty urns, has the asymptotic form*

$$p_n(r) \sim p^*(r) + \delta_r(\log n), \quad n \to \infty;$$

where $\delta_r(x)$ is a periodic function, with 0 mean.

2.2.2 Asymptotic expansions for $p \to 1$ and $p \to 0$.

Some asymptotic expansions for $p^*(1)$ are computed as follows, as $p \to 1$ and $p \to 0$.

1. For $p = 1 - \varepsilon$, $\varepsilon \to 0$, we derive for the first term in the summation in (19) ($u = 1$)

$$\begin{aligned}\left[\ln\left(1 + 1/(\alpha + Q)\right) - \ln\left(1 + 1/(\alpha + Q + Q^2)\right)\right]/L& \\ = -\varepsilon/\ln(\varepsilon) + 3\varepsilon^2/(2\ln(\varepsilon) - \varepsilon^3/(3\ln(\varepsilon)) + \cdots .&\end{aligned}$$

   The next term ($u = 2$) gives

$$\left[\ln\left(1 + 1/(\alpha + Q^2)\right) - \ln\left(1 + 1/(\alpha + Q + Q^2)\right)\right]/L = -\varepsilon^3/\ln(\varepsilon) + \cdots .$$

   These first two terms give

$$p^*(1) \sim -\varepsilon/\ln(\varepsilon) + 3\varepsilon^2/(2\ln(\varepsilon)) - 4\varepsilon^3/(3\ln(\varepsilon)) + \cdots . \tag{20}$$

2. For $p = \varepsilon$, $\varepsilon \to 0$, the saddle point analysis (as in a previous section) still applies, but (19) introduces a sum

$$\sum_{u=1}^\infty \frac{e^{-yQ^u}}{1 - e^{-yQ^u}} \sim \sum_{u=1}^\infty \frac{e^{-ye^{\varepsilon u}}}{1 - e^{-ye^{\varepsilon u}}} \sim \int_1^\infty \frac{e^{-yt}}{1 - e^{-yt}} \frac{dt}{\varepsilon t} + \frac{1}{2} \frac{e^{-y}}{1 - e^{-y}}$$

   by Euler-Maclaurin.

   Set

$$K_1 := \int_1^\infty \frac{e^{-\tilde{y}t}}{1 - e^{-\tilde{y}t}} \frac{dt}{t} = 0.57149987436\ldots .$$

   This gives $p^*(1) \sim p^*(0)K_1/\varepsilon$ where $p^*(0)$ is given by (12) ($p^*(0) \to 0$, exponentially, $\varepsilon \to 0$) and more generally, $p^*(k) \sim p^*(0)K_k/\varepsilon^k$ for fixed k, where K_k is a numerical integral. For instance,

$$K_2 = \int_1^\infty \frac{e^{-\tilde{y}t_1}}{1 - e^{-\tilde{y}t_1}} \int_{t_1}^\infty \frac{e^{-\tilde{y}t_2}}{1 - e^{-\tilde{y}t_2}} \frac{dt_2}{t_2} \frac{dt_1}{t_1} = 0.1633054070\ldots$$

   Asymptotics of $p^*(k)$ for large k (k of order $-\ln(\varepsilon)/\varepsilon$ for instance) will be considered in Section 4.

2.2.3 Simplifications for $p = 1/2$

Let us first consider $p_n(0)$. We can think of the probabilities in terms of coin flippings: Each individual flips a coin, and if the outcome is "0", the item finds its place in urn 1; the remaining individuals continue flipping a coin, falling into urn 2 if they have flipped "0", otherwise they continue, and so on. That leads to an exact recursion (this one was already mentioned in [11]):

$$p_n(0) = 2^{-n}\sum_{u=1}^{n}\binom{n}{u}p_{n-u}(0), \quad n \geq 1,\ p_0(0) = 1.$$

In this sum, $u \geq 1$, since at least one ball must go into urn 1 (gap-free!).

Induction proves immediately $p_n(0) = \frac{1}{2}$.

Now, let $p_n(1)$ be the probability of 1 empty urn. The recursion is:

$$p_n(1) = 2^{-n}\sum_{u=1}^{n}\binom{n}{u}p_{n-u}(1) + 2^{-n}p_n(0), \quad n \geq 1,\ p_0(1) = 0.$$

Induction proves immediately $p_n(1) = \frac{1}{4}$.

Now consider the general situation of r empty urns. The recursion is:

$$p_n(r) = 2^{-n}\sum_{u=1}^{n}\binom{n}{u}p_{n-u}(r) + 2^{-n}p_n(r-1), \quad n \geq 1,\ p_0(r) = [\![r=0]\!].$$

Induction proves immediately

$$p_n(r) = \frac{1}{2^{r+1}}. \tag{21}$$

Note that the generating function $F^*(z)$ of $p^*(r) : F^*(z) := \mathbb{E}^*(z^X)$ is given by

$$F^*(z) = \frac{1}{2-z}, \tag{22}$$

which immediately gives all moments.

For the instance $r = 0$ (gap-free) that argument leads to a nice "bijection": We consider the empty urn with lowest index. (In terms of the coin flippings this mean that all individuals have flipped 1 in the corresponding round.) Now we move all balls in urns with a higher index into this empty urn. This preserves probability and transforms "no empty urn" $\longleftrightarrow$ "empty urns". (The inverse mapping is to take the items in the last nonempty urn and distribute them at random in urns with higher indices.)

Let us now consider our previous asymptotic expressions. When $p = 1/2$, we can again simplify (18):

$$\varphi(1,s) := \int_0^\infty y^{s-1}e^{-y}\left[\sum_{j=0}^{\infty}(-1)^{\nu(j)}e^{-jy}\right]\sum_{u=1}^{\infty}\frac{e^{-y2^u}}{1-e^{-y2^u}}dy.$$

We will not use this expression in the sequel, but instead, we first turn to $p^*(0)$ and, similarly to (9), let

$$f(\beta) = \frac{1}{L}\int_0^\infty e^{-\beta t}\prod_{j\geq 0}\left(1-e^{-2^j t}\right)\frac{dt}{t}.$$

We want to show (independently from [3]) that $f(1) = p^*(0) = \frac{1}{2}$. We derive the functional equation

$$f(\beta) = f(\tfrac{\beta}{2}) - f(\tfrac{\beta+1}{2})$$

by substituting $t \to t/2$ in the integral and rearranging. Plugging in $\beta = 1$ leads to the equation $f(1) = \frac{1}{2} f(\frac{1}{2})$. Now

$$f(\tfrac{1}{2}) = \lim_{\beta \to 0} \Big(f(\tfrac{\beta}{2}) - f(\beta) \Big).$$

The only term that remains when evaluating this limit is

$$\lim_{\beta \to 0} \frac{1}{L} \int_0^\infty \frac{e^{-\beta t/2} - e^{-\beta t}}{t} dt.$$

But this integral is independent of β and evaluates to 1, as required.

The same simplifications can be obtained for $p^*(1)$ and, more generally, for $p^*(r)$: see the details in the long version of this paper [18].

2.2.4 An inclusion-exclusion argument for $p = 1/2$

Now we use an inclusion-exclusion argument to describe $p_n(0)$ and, more generally, $p_n(r)$. Since we know *a priori* that the outcome of this computation must be $1/2^{r+1}$, this approach will lead us to some unexpected and exciting identities.

Let, as usual k denote the last full urn. Consider the following random variables (we use Iverson's notation $[\![P]\!]$, which is 1 if P is true, and 0 otherwise)

$$V_i := [\![\text{urn } i \text{ empty}]\!], \quad \overline{V_{k+1}} := [\![\text{all urns from } k+1 \text{ on empty}]\!].$$

It is clear that $k > n$ is impossible: n balls cannot go into $> n$ urns, which are all non-empty. Then, exactly,

$$\begin{aligned} p_n(0) &= \mathbb{E}\left\{ \sum_{k=1}^{n} (1 - V_1) \dots (1 - V_k) \overline{V_{k+1}} \right\} \\ &= \mathbb{E}\left\{ \sum_{k=1}^{n} \left[1 - \sum_{i=1}^{k} V_i + \sum_{1 \le i < j \le k} V_i V_j + \dots + (-1)^{k-1} \sum_{1 \le i_1 < i_2 < \dots < i_{k-1} \le k} V_{i_1} V_{i_2} \dots V_{i_{k-1}} \right] \overline{V_{k+1}} \right\} \\ &= \sum_{k=1}^{n} \sum_{r=0}^{k-1} (-1)^r \sum_{1 \le i_1 < i_2 < \dots < i_r \le k} \left[1 - \sum_{s=1}^{r} \frac{1}{2^{i_s}} - \frac{1}{2^k} \right]^n . \end{aligned} \tag{23}$$

This equals 1/2, since this value is known from [11]. More generally

$$\begin{aligned} p_n(g) &= \mathbb{E}\Big\{ \sum_{k=1+g}^{n+g} \Big[\sum_{1 \le u_1 < u_2 \dots < u_g < k} \sum_{r=0}^{k-1-g} \sum_{1 \le i_1 < i_2 < \dots < i_r \le k} (-1)^r [\![u_l \ne i_s]\!] V_{i_1} \dots V_{i_r} V_{u_1} \dots V_{u_g} \Big] \overline{V_{k+1}} \Big\} \\ &= \sum_{k=1+g}^{n+g} \sum_{1 \le u_1 < u_2 \dots < u_g < k} \sum_{r=0}^{k-1-g} \sum_{1 \le i_1 < i_2 < \dots < i_r \le k} (-1)^r [\![u_l \ne i_s]\!] \left[1 - \sum_{s=1}^{r} \frac{1}{2^{i_s}} - \sum_{l=1}^{g} \frac{1}{2^{u_l}} - \frac{1}{2^k} \right]^n . \end{aligned} \tag{24}$$

This should give $1/2^{g+1}$, as we will see now.

Of course, equations (23) and (24) can be generalized to $p \neq 1/2$, but they do not lead to the following simplifications.

The inner sum in equation (23) leads to

$$\begin{aligned}
&\sum_{I\subseteq\{1,\dots,k\}} (-1)^{|I|}\left[1-\sum_{i\in I}\frac{1}{2^i}-\frac{1}{2^k}\right]^n \\
&= \sum_{J\subseteq\{0,\dots,k-1\}} (-1)^{|J|}\left[1-\frac{1}{2^k}\sum_{i\in J}2^i-\frac{1}{2^k}\right]^n.
\end{aligned}$$

This can be elegantly written using the function $\nu(j)$. Therefore we obtain the identity

$$\sum_{k=1}^{n}\sum_{0\le\lambda<2^k}(-1)^{\nu(\lambda)}\left(1-\frac{1+\lambda}{2^k}\right)^n=\frac{1}{2}. \tag{25}$$

It can also be written as

$$\sum_{k\ge1}\sum_{0\le\lambda<2^k}(-1)^{\nu(\lambda)}\left(1-\frac{1+\lambda}{2^k}\right)^n=\frac{1}{2},$$

since for $k > n$ no new terms are added. (The previous computation gives zero for values k that are outside the natural range.)

Let us now consider the case of one empty urn. The inner sum in (24) is

$$\begin{aligned}
&\sum_{1\le u<k}\ \sum_{I\subseteq\{1,\dots,k\}\setminus\{u\}} (-1)^{|I|}\left[1-\sum_{i\in I}\frac{1}{2^i}-\frac{1}{2^u}-\frac{1}{2^k}\right]^n \\
&= \sum_{1\le u\le k}\ \sum_{I\subseteq\{1,\dots,k\}\setminus\{u\}} (-1)^{|I|}\left[1-\sum_{i\in I}\frac{1}{2^i}-\frac{1}{2^u}-\frac{1}{2^k}\right]^n - \sum_{I\subseteq\{1,\dots,k-1\}} (-1)^{|I|}\left[1-\sum_{i\in I}\frac{1}{2^i}-\frac{1}{2^k}-\frac{1}{2^k}\right]^n \\
&=: A-B.
\end{aligned}$$

Now

$$\begin{aligned}
A &= \sum_{1\le u\le k}\ \sum_{I\subseteq\{1,\dots,k\}\setminus\{u\}} (-1)^{|I|}\left[1-\sum_{i\in I}\frac{1}{2^i}-\frac{1}{2^u}-\frac{1}{2^k}\right]^n \\
&= \sum_{J\subseteq\{1,\dots,k\}} (-1)^{|J|+1}|J|\left[1-\sum_{i\in J}\frac{1}{2^i}-\frac{1}{2^k}\right]^n \\
&= \sum_{J\subseteq\{0,\dots,k-1\}} (-1)^{|J|+1}|J|\left[1-\frac{1}{2^k}\sum_{i\in J}2^i-\frac{1}{2^k}\right]^n \\
&= \sum_{0\le\lambda<2^k} (-1)^{\nu(\lambda)+1}\nu(\lambda)\left[1-\frac{1+\lambda}{2^k}\right]^n.
\end{aligned}$$

Likewise,

$$
\begin{aligned}
B &= \sum_{I\subseteq\{1,\dots,k-1\}} (-1)^{|I|}\left[1-\sum_{i\in I}\frac{1}{2^i}-\frac{1}{2^{k-1}}\right]^n \\
&= \sum_{I\subseteq\{0,\dots,k-2\}} (-1)^{|I|}\left[1-\frac{1}{2^{k-1}}\sum_{i\in I}2^i-\frac{1}{2^{k-1}}\right]^n \\
&= \sum_{0\le\lambda<2^{k-1}} (-1)^{\nu(\lambda)}\left[1-\frac{1+\lambda}{2^{k-1}}\right]^n.
\end{aligned}
$$

Summarizing

$$
p_n(1)=\sum_{k=2}^{n+1}\left[\sum_{0\le\lambda<2^k}(-1)^{\nu(\lambda)+1}\nu(\lambda)\Big(1-\frac{1+\lambda}{2^k}\Big)^n-\sum_{0\le\lambda<2^{k-1}}(-1)^{\nu(\lambda)}\Big(1-\frac{1+\lambda}{2^{k-1}}\Big)^n\right]=\frac{1}{4}.
$$

In general, for r empty urns, we obtain with an analogous argument

$$
p_n(r)=\sum_{k=r+1}^{n+r}\left[\sum_{0\le\lambda<2^k}(-1)^{\nu(\lambda)+r}\binom{\nu(\lambda)}{r}\Big(1-\frac{1+\lambda}{2^k}\Big)^n+\sum_{0\le\lambda<2^{k-1}}(-1)^{\nu(\lambda)+r}\binom{\nu(\lambda)}{r-1}\Big(1-\frac{1+\lambda}{2^{k-1}}\Big)^n\right]
$$

(for $r=0$, the second term isn't there).

We can alternatively sum over $k\ge 1$. But then we get by a simple rearrangement the beautiful formula

$$
\begin{aligned}
p_n(r) &= \sum_{k\ge1}\sum_{0\le\lambda<2^k}(-1)^{\nu(\lambda)+r}\binom{\nu(\lambda)}{r}\Big(1-\frac{1+\lambda}{2^k}\Big)^n+\sum_{k\ge1}\sum_{0\le\lambda<2^k}(-1)^{\nu(\lambda)+g}\binom{\nu(\lambda)}{r-1}\Big(1-\frac{1+\lambda}{2^k}\Big)^n \\
&= \sum_{k\ge1}\sum_{0\le\lambda<2^k}(-1)^{\nu(\lambda)+r}\binom{\nu(\lambda)+1}{r}\Big(1-\frac{1+\lambda}{2^k}\Big)^n=\frac{1}{2^{r+1}}. \qquad (26)
\end{aligned}
$$

3 The moments of the total number of empty urns

Now we turn our attention from the probabilities of a certain number of empty urns to the total number X of empty urns and the moments of this random variable. First we analyze the mean of X. Next we consider the case $p\to1$ and $p\to0$. Then we obtain the asymptotic distribution of X, when $p\to0$ and finally consider the case $p=1/2$.

3.1 The general case

We have, with Theorem 2.1,

$$
\mathbb{E}(e^{vX})\sim\sum_{k=1}^{\infty}\left[1-e^{-n^*/Q^k}\right]e^{-\alpha n^*/Q^k}\prod_{w=1}^{\infty}\left[(e^v-1)e^{-n^*Q^w/Q^k}+1\right],\quad n\to\infty
$$

which leads to

$$
\Lambda(v,s):=\int_0^\infty y^{s-1}e^{-\alpha y}(1-e^{-y})\exp\left[\sum_{i=1}^{\infty}(-1)^{i+1}\frac{(e^v-1)^i}{i}V_i(y)\right]dy,
$$

with

$$V_j(y) := \sum_{w=1}^{\infty} e^{-jye^{Lw}}.$$

We derive

$$\Lambda(v,s) = \int_0^{\infty} y^{s-1} e^{-\alpha y}\left(1-e^{-y}\right)\left[1 + vV_1 + \frac{v^2}{2}[V_1 - V_2 + V_1^2] + \cdots\right] dy. \tag{27}$$

Expanding, the dominant part $\mathbb{E}^*(X)$ of $\mathbb{E}(X)$ is given by

$$\mathbb{E}^*(X) = \sum_{w=1}^{\infty} \int_0^{\infty} e^{-\alpha y}\left[\sum_{u=1}^{\infty} \frac{(-1)^{u+1} y^u}{u!}\right] e^{-yQ^w} \frac{dy}{Ly} = \frac{1}{L}\sum_{w=1}^{\infty} \ln\left(1 + \frac{1}{\alpha + Q^w}\right). \tag{28}$$

Next we set

$$\Lambda_1(s) := \sum_{w=1}^{\infty} \int_0^{\infty} y^{s-1} e^{-\alpha y}\left[1 - e^{-y}\right] e^{-yQ^w} dy.$$

So,

$$\Lambda_1(s) = \Gamma(s) \sum_{w \geq 1}\left[\frac{1}{(\alpha + Q^w)^s} - \frac{1}{(1 + \alpha + Q^w)^s}\right]. \tag{29}$$

The periodic component $\mathbb{E}_P(X)$ of $\mathbb{E}(X)$ is given by

$$\mathbb{E}_P(X) = \frac{1}{L}\sum_{l \neq 0} \Lambda_1(\chi_l) e^{-2l\pi \mathbf{i} \log n^*}.$$

This can be simplified as follows. Note that

$$\lim_{w \to \infty} \frac{1}{(\alpha + Q^w)^{\chi_l}} = 1.$$

So

$$\Lambda_1(\chi_l) = \Gamma(\chi_l) \sum_{w \geq 1}\left[\left(\frac{1}{(\alpha + Q^w)^{\chi_l}} - 1\right) - \left(\frac{1}{(1 + \alpha + Q^w)^{\chi_l}} - 1\right)\right].$$

But

$$\frac{1}{(1 + \alpha + Q^{w+1})^{\chi_l}} = \frac{1}{((1+\alpha)/Q + Q^w)^{\chi_l}} = \frac{1}{(\alpha + Q^w)^{\chi_l}}.$$

By telescoping,

$$\Lambda_1(\chi_l) = \Gamma(\chi_l)\left[1 - \frac{1}{(1 + \alpha + Q)^{\chi_l}}\right] = \Gamma(\chi_l)[1 - p^{\chi_l}]. \tag{30}$$

Note carefully, that this simplification applies to all values of p, with $p = 1/2$ being particularly simple, since then $1 - p^{\chi_l} = 0$, and the Fourier coefficients disappear.

To summarize, we have this theorem.

Theorem 3.1 *The dominant part $\mathbb{E}^*(X)$ of $\mathbb{E}(X)$ is given by*

$$\mathbb{E}^*(X) = \frac{1}{L}\sum_{w=1}^{\infty} \ln\left(1 + \frac{1}{\alpha + Q^w}\right).$$

The periodic component of $\mathbb{E}_P(X)$ of $\mathbb{E}(X)$ is given by

$$\mathbb{E}_P(X) = \frac{1}{L}\sum_{l\neq 0} \Lambda_1(\chi_l) e^{-2l\pi\mathbf{i}\log n^*},$$

with

$$\Lambda_1(\chi_l) = \Gamma(\chi_l)[1 - p^{\chi_l}].$$

Of course the other moments can be expressed almost automatically by similar (more complicated) formulæ.

The generating function $F_n(z)$ of the $p_n(k)$'s, i.e., $F_n(z) := \mathbb{E}(z^X)$ is asymptotically given, with Theorem 2.1, by

$$\sum_{k=1}^{\infty}\left[1 - e^{-n^*/Q^k}\right] e^{-\alpha n^*/Q^k} \prod_{w=1}^{\infty}\left[(z-1)e^{-n^* Q^w/Q^k} + 1\right],$$

which leads to

$$\Lambda(z,s) := \int_0^{\infty} y^{s-1} e^{-\alpha y}(1 - e^{-y}) \exp\left[\sum_{i=1}^{\infty}(-1)^{i+1}\frac{(z-1)^i}{i} V_i(y)\right] dy, \tag{31}$$

with

$$V_j(y) := \sum_{w=1}^{\infty} e^{-jye^{Lw}}.$$

3.1.1 Asymptotic expressions for $p \to 1$ and $p \to 0$

Some asymptotic expressions are computed as follows, for $p \to 1$ and $p \to 0$.

1. For $p = 1 - \varepsilon$, $\varepsilon \to 0$, the dominant part of the mean gives, from (28),

$$\mathbb{E}^*(X) \sim -1/\ln(\varepsilon)\sum_{w=1}^{\infty}\ln\left(1 + \frac{1}{\varepsilon + (1/\varepsilon)^w}\right) \sim -1/\ln(\varepsilon)\sum_{w=1}^{\infty}\varepsilon^w \sim \frac{-\varepsilon}{\ln(\varepsilon)}.$$

   Note that this corresponds, to first order, to $p^*(1)$ as given by (20).

2. For $p = \varepsilon$, $\varepsilon \to 0$, the dominant part of the mean gives, again from (28),

$$\begin{aligned}\mathbb{E}^*(X) &\sim 1/\varepsilon\sum_{w=1}^{\infty}\ln\left(1 + \frac{1}{1/\varepsilon + e^{\varepsilon w}}\right)\\ &\sim 1/\varepsilon\int_1^{\infty}\frac{1}{1/\varepsilon + v}\frac{dv}{v\varepsilon} + \mathcal{O}(\ln(\varepsilon))\\ &= 1/\varepsilon\int_{\varepsilon}^{\infty}\frac{du}{u(1+u)} + \mathcal{O}(\ln(\varepsilon)) \sim -\ln(\varepsilon)/\varepsilon + \mathcal{O}(\ln(\varepsilon)).\end{aligned} \tag{32}$$

   For $\varepsilon = 0.01$, (28) gives $457.7862279\ldots$ and the asymptotic expression gives $460\ldots$.

We now turn to the dominant term of the variance: for $p = \varepsilon$, this is given by $\mathrm{VAR}^*(X) = \mathbb{E}^*(X^2) - [\mathbb{E}^*(X)]^2$. First we must compute the mean with more precision. We start from (28) and consider each source of errors. The first type of error leads to $\mathcal{O}(\ln(\varepsilon))$: it is due to

- replacing L by ε,
- replacing α by $1/\varepsilon$,
- replacing Q^w by $e^{\varepsilon w}$,
- using only the first term in the expansion of $\ln\left(1 + \frac{1}{1/\varepsilon + e^{\varepsilon w}}\right)$, the next term leads to $-1/2 \int_{v=1}^{\infty} \frac{1}{(1+\varepsilon v)^2} \frac{dv}{v}$.

The second type of error leads to $\mathcal{O}(1)$: it is due to

- using Euler-Mc Laurin,
- starting the integral in (32) at $v = 1$ instead of $v = e^{\varepsilon}$,
- using only the first term in the expansion of $\int_{u=\varepsilon}^{\infty} \frac{du}{u(1+u)} = \ln(1+\varepsilon) - \ln(\varepsilon) \sim -\ln(\varepsilon) + \varepsilon$.

So, finally,

$$\mathbb{E}^*(X) \sim -\ln(\varepsilon)/\varepsilon + \mathcal{O}(\ln(\varepsilon)), \quad \varepsilon \to 0.$$

The second term in the variance is thus given by

$$[\mathbb{E}^*(X)]^2 \sim \ln^2(\varepsilon)/\varepsilon^2 + \mathcal{O}(\ln^2(\varepsilon)/\varepsilon), \quad \varepsilon \to 0.$$

The first term in the variance is related, from (27), to $V_1 - V_2 + V_1^2$, which leads to

$$\frac{1}{L}\sum_{w=1}^{\infty} \ln\left(1 + \frac{1}{\alpha + Q^w}\right) - \frac{1}{L}\sum_{w=1}^{\infty} \ln\left(1 + \frac{1}{\alpha + 2Q^w}\right) + \frac{1}{L}\sum_{w_1=1}^{\infty}\sum_{w_2=1}^{\infty} \ln\left(1 + \frac{1}{\alpha + Q^{w_1} + Q^{w_2}}\right). \tag{33}$$

When $p = \varepsilon$, the first sum leads to

$$-\ln(\varepsilon)/\varepsilon + \mathcal{O}(\ln(\varepsilon)).$$

The second sum leads to

$$\ln(\varepsilon)/\varepsilon + \mathcal{O}(\ln(\varepsilon)).$$

The third sum (which is the most important) leads asymptotically to

$$1/\varepsilon \int_{v_1=1}^{\infty}\int_{v_2=1}^{\infty} \frac{1}{1/\varepsilon + v_1 + v_2} \frac{dv_1}{v_1 \varepsilon} \frac{dv_2}{v_2 \varepsilon} \sim \frac{\ln^2(\varepsilon) + \pi^2/6}{\varepsilon^2} + \mathcal{O}(\ln^2(\varepsilon)/\varepsilon).$$

We finally obtain

$$\mathrm{VAR}^*(X) \sim \frac{\pi^2}{6\varepsilon^2}, \quad \varepsilon \to 0.$$

The other moments can be derived similarly, but we will provide a better way in Section 3.2.

Another interesting problem is the analysis of periodicities in the mean, when $p = \varepsilon$, $\varepsilon \to 0$. We know from (30) that this is asymptotically given, with $\chi_l \sim 2\pi i l/\varepsilon$ by

$$\begin{aligned}
&\sum_{l\neq 0} \frac{\Gamma(2\pi\mathbf{i}l/\varepsilon)}{\varepsilon}[1-\varepsilon^{\chi_l}]e^{-2l\pi\mathbf{i}\ln(n\varepsilon)/\varepsilon} \\
\sim\ &\sum_{l\neq 0} \frac{\Gamma(2\pi\mathbf{i}l/\varepsilon)}{\varepsilon}[\varepsilon^{-\chi_l}-1]e^{-2l\pi\mathbf{i}\ln(n)/\varepsilon}. \qquad (34)
\end{aligned}$$

This goes to 0 as ε goes to 0, due to the rapid decrease of Γ towards $\mathbf{i}\infty$.

3.2 The asymptotics for $p^*(r)$, $p = \varepsilon$, $\varepsilon \to 0$

Inspired by Sections 2.1 and 3, we write the asymptotic generating function $F^*(z)$ of the $p^*(r)$'s: $F^*(z) := \mathbb{E}^*(z^X)$ as

$$\begin{aligned}
F^*(z) &\sim \int_0^\infty e^{-y/\varepsilon}\exp\left\{\int_1^\infty \ln\left[1+(z-1)e^{-ye^{\varepsilon l}}\right]dl\right\}(1-e^{-y})\frac{dy}{yL} \\
&\sim \int_0^\infty e^{-y/\varepsilon}\exp\left\{\frac{1}{\varepsilon}\int_1^\infty \ln\left[1+(z-1)e^{-yv}\right]\frac{dv}{v}\right\}(1-e^{-y})\frac{dy}{yL}, \quad \varepsilon\to 0,
\end{aligned}$$

and we have

$$p^*(r) = \frac{1}{2\pi\mathbf{i}}\int_\Omega F^*(z)\frac{dz}{z^{r+1}},$$

where Ω is inside the analyticity domain of the integrand and encircles the origin. We set $z = 1+u\varepsilon$. (This choice will be justified later on.) We derive

$$\begin{aligned}
\frac{1}{\varepsilon}\int_1^\infty \ln\left[1+(z-1)e^{-yv}\right]\frac{dv}{v} &\sim \frac{1}{\varepsilon}\int_1^\infty e^{-yv}u\varepsilon\frac{dv}{v} \sim u\int_y^\infty e^{-t}\frac{dt}{t} = u\mathrm{Ei}(1,y) \\
&\sim -u\ln(y)+u[-\gamma+\mathcal{O}(y)] \quad \text{for } y \text{ small.}
\end{aligned}$$

Setting $y = \varepsilon w$, we obtain

$$\begin{aligned}
F^*(z) &\sim \frac{1}{L}\int_0^\infty e^{-w}e^{-u[\ln(w)+\ln(\varepsilon)+\gamma]}\frac{(1-e^{-\varepsilon w})dw}{w} \sim \frac{\Gamma(-u)}{\varepsilon}\left[1-(1+\varepsilon)^u\right]e^{-u\gamma}e^{-u\ln(\varepsilon)} \\
&\sim \Gamma(1-u)e^{-u\gamma}e^{-u\ln(\varepsilon)}, \quad \varepsilon\to 0.
\end{aligned}$$

Next we derive

$$\mathbb{E}^*\left[e^{u\varepsilon X}\right] \sim \mathbb{E}^*\left[z^X\right] \sim \Gamma(1-u)e^{-u\gamma}e^{-u\ln(\varepsilon)}, \quad \varepsilon\to 0,$$

justifying the choice $z = 1+u\varepsilon$.

Setting $\mu := -\ln(\varepsilon)/\varepsilon$ (this is the dominant term of the mean), we obtain

$$\mathbb{E}^*\left[e^{u\varepsilon(X-\mu)}\right] \sim \Gamma(1-u)e^{-u\gamma}, \quad \varepsilon\to 0.$$

From this we get

Proposition 3.2 *The dominant part of*

$$\varepsilon(X-\mu)+\gamma, \quad \varepsilon\to 0,$$

is asymptotically distributed as a Gumbel extreme-value random variable.

Of course we immediately recover the variance previously computed. All asymptotic moments are now easily obtained.

3.3 Simplifications for $p = 1/2$

Again, we can show independently, that our asymptotic expressions are consistent with (22). The interested reader can find the details of these computations in the long version of this paper [18].

Here, we only cite a curious formula that we obtained as a byproduct:

$$\frac{1}{L} \ln \prod_{j \geq 0} \left(\frac{2j+2}{2j+1} \right)^{(z-1)^{\nu(j)}} = \frac{1}{2-z}.$$

Of course, letting $z \to 0$, we recover (16).

4 The distribution of the last full urn K, conditioned on the number of empty urns

What is the conditional probability $p_n(0,k)$ that the last full urn K is urn k, given that there are no empty urns, for $p = 1/2$? From (6), this is asymptotically given, with $\eta = k - \log n$, by

$$p_n(0,k) \sim f(\eta) = 2 \prod_{u=0}^{\infty} \left[1 - \exp\left(e^{-L(\eta-u)}\right)\right] \exp\left(e^{-L(\eta)}\right), \quad n \to \infty.$$

We have analyzed in detail, in [17, Section 5.9], a similar distribution: the distribution of the first empty urn. This corresponds also to the Probabilistic Counting (see Flajolet and Martin [7]) and to the first empty part in compositions of integers (see [12]).

We have here another example of Gumbel-like distribution, the distribution function of which is $\exp(-e^{-x})$. The rate of convergence for this kind of distributions is fully analyzed in [17]; we will not give the details here. Let $P_n(0,k) := \sum_{i=0}^{k} p_n(0,i)$. We have $P_n(0,k) \sim F(\eta)$, with $f(\eta) = F(\eta) - F(\eta - 1)$. We compute

$$\phi(\alpha) = \int_{-\infty}^{\infty} e^{\alpha\eta} f(\eta) d\eta = 2M(-\widetilde{\alpha})\Gamma(-\widetilde{\alpha})/L,$$

where $M(s)$ is defined in (15) and $\widetilde{\alpha} := \alpha/L$. This is similar to (13). We have $\phi(0) = 1$, hence $M(0) = 0$ (which is also derived from $N(0) = -1$) and $M'(0) = L/2$ (which is also derived from (16)). Also $M(\chi_l) = 0$ from (17). We will need $M''(0)$, $M'''(0)$, ... we can for instance proceed as in [17]:

$$\begin{aligned} M(s) = \sum_{j=0}^{\infty} \frac{(-1)^{\nu(j)}}{(8j+1)^s} \Bigg[& 1 - \frac{1}{(1+1/(8j+1))^s} - \frac{1}{(1+2/(8j+1))^s} + \frac{1}{(1+3/(8j+1))^s} \\ & - \frac{1}{(1+4/(8j+1))^s} + \frac{1}{(1+5/(8j+1))^s} + \frac{1}{(1+6/(8j+1))^s} - \frac{1}{(1+7/(8j+1))^s} \Bigg]. \end{aligned} \tag{35}$$

This leads to

$$\begin{aligned} M(0) &= 0, \\ M'(0) &= L/2, \\ M''(0) &= .3378750212\ldots, \\ M'''(0) &= -.9600749346\ldots. \end{aligned}$$

Still another family of identities can be obtained. For instance, from (35), we derive

$$M'(0) = L/2 = \sum_{j=0}^{\infty} (-1)^{\nu(j)} \ln \frac{2(4j+1)(8j+3)(8j+5)(j+1)}{(8j+1)(2j+1)(4j+3)(8j+7)}.$$

(This would also follow from the identity (16)).

Now, we can compute (almost) mechanically all moments of $K - \log n$ we need, from $\phi(\alpha)$, using techniques and notations from [17]. Let m_i be the ith moment of $F(\eta)$, $\tilde{m}_1$ be the dominant part of the ith moment of $K - \log n$, w_i be the corresponding periodic part, μ_i be the ith centered moment of $F(\eta)$, $\tilde{\mu}_1$ be the dominant part of the ith centered moment of $K - \log n$, κ_i be the corresponding periodic part. We have

Theorem 4.1 *The moments of $K - \log n$ are given by*

$$\begin{aligned}
m_1 &= -M''(0)/L^2 + \gamma/L - 1/2, \\
m_2 &= (-2M''(0)\gamma + 2M'''(0)/3)/L^3 + (\pi^2/6 + M''(0) + \gamma^2)/L^2 - \gamma/L + 1/6, \\
\tilde{m}_1 &= -M''(0)/L^2 + \gamma/L, \\
\tilde{m}_2 &= (-2M''(0)\gamma + 2M'''(0)/3)/L^3 + (\pi^2/6 + \gamma^2)/L^2, \\
\mu_2 &= -M''(0)^2/L^4 + 2M'''(0)/(2L^3) + \pi^2/(6L^2) - 1/12, \\
\tilde{\mu}_2 &= -M''(0)^2/L^4 + 2M'''(0)/(3L^3) + \pi^2/(6L^2).
\end{aligned}$$

Let w_1 denote the periodic component of the mean and κ_2 denote the periodic component of the variance, then

$$\begin{aligned}
w_1 &= \sum_{l \neq 0} -2M'(\chi_l)\Gamma(\chi_l) e^{-2l\pi\mathbf{i}\log n}/L^2, \\
\kappa_2 &= \sum_{l \neq 0} \Big[-4M'(\chi_l)M''(0)\Gamma(\chi_l)/L^4 \\
&\quad + 2\left(2M'(\chi_l)\gamma + 2M'(\chi_l)\psi(\chi_l) + M''(\chi_l)\right)\Gamma(\chi_l)/L^3 \Big] e^{-2l\pi\mathbf{i}\log n}/L^2 - w_1^2.
\end{aligned}$$

Note that we could also have started from (25). For instance (26) leads to the asymptotic distribution of K given r empty urns, with some generalized $M_i(s)$. We omit the details.

5 Super gap-free

Let $p(n)$ be the probability that a random sample of size n leads to a gap-free situation, where each non-empty urn contains at least 2 balls. We confine ourselves to $p = q = 1/2$ here. The recursion

$$p(n) = 2^{-n} \sum_{j=2}^{n} \binom{n}{j} p(n-j), \quad n \geq 2, \quad p(0) = 1, \; p(1) = 0$$

is easy to understand, if one thinks about coin flippings: at least two out of the n "players" must flip a zero and thus go into the first urn. In terms of the exponential generating function

$$F(z) := \sum_{n \geq 0} p(n) \frac{z^n}{n!},$$

this means

$$F(z) = 1 + (e^{z/2} - 1 - z/2)F(z/2).$$

Now we set

$$\varphi(z) := e^{-z}F(z) = \sum_{n\ge 0} b(n)\frac{z^n}{n!}, \quad b(0) = 1,\ b(1) = -1.$$

Depoissonization (as described in the book by Szpankowski [23]) leads to $p(n) \sim \varphi(n)$. In order to find $\varphi(z)$, for $z \to \infty$, we first note that

$$\varphi(z) = \varphi(z/2) + R(z),$$

with

$$R(z) = e^{-z} - e^{-z}\big(1 + \tfrac{z}{2}\big)F(z/2).$$

For technical reasons (Mellin transform must exist), we set

$$\varphi_1(z) := \varphi(z) - 1 = -z + \cdots,$$

so that $p(n) \sim 1 + \varphi_1(n)$.

The Mellin transform is

$$\varphi_1^*(s) = 2^s\varphi_1^*(s) + R^*(s) = \frac{R^*(s)}{1 - 2^s},$$

and so

$$\varphi_1(z) = \frac{1}{2\pi\mathbf{i}}\int_{-1/2-\mathbf{i}\infty}^{-1/2+\mathbf{i}\infty} \frac{R^*(s)}{1-2^s} z^{-s}ds.$$

From this, we get in the usual way, by computing residues,

$$\varphi_1(z) = R^*(0)/L + \frac{1}{L}\sum_{l\ne 0} R^*(\chi_l)e^{-2l\pi\mathbf{i}\log z} + \mathcal{O}(z^{-C}),$$

for some $C > 0$. Now

$$R(z) = e^{-z} - e^{-z}\big(1 + \tfrac{z}{2}\big)\sum_{j\ge 0}\frac{p(j)z^j}{2^j j!},$$

so

$$R^*(s) = \Gamma(s) - \sum_{j\ge 0}\frac{p(j)}{2^j j!}\Big[\Gamma(s+j) + \frac{1}{2}\Gamma(s+j+1)\Big].$$

Hence

$$\begin{aligned} R^*(0)/L &= -\frac{1}{L}\sum_{j\ge 1}\frac{p(j)}{2^j j!}\Gamma(j) - \frac{1}{L}\sum_{j\ge 0}\frac{p(j)}{2^j j!}\frac{\Gamma(j+1)}{2} \\ &= -\frac{1}{L}\sum_{j\ge 1}\frac{p(j)}{2^j}\Big(\frac{1}{j} + \frac{1}{2}\Big) - \frac{1}{2L} = -0.8499303988\ldots \end{aligned}$$

From this, we get (apart from fluctuations)

$$p(n) \sim 1 + R^*(0)/L = 0.150069011\ldots.$$

We compute directly $p(400) = 0.1500696051\ldots$ and "see" from that that the periodic term is of order 10^{-6}.

Let us now quickly sketch the situation of d-super-gap-free, for which we get the recursion

$$p(n) = 2^{-n}\sum_{j=d}^{n}\binom{n}{j}p(n-j), \quad n \ge 2, \quad p(0)=1,\; p(1)=0.$$

This leads to

$$F(z) = 1 + \Big(e^{z/2} - e_{d-1}(\tfrac{z}{2})\Big)F(z/2),$$

with the truncated exponential series $e_m(z) = 1 + z + \cdots + \frac{z^m}{m!}$.

Again, we find

$$\varphi(z) = \varphi(z/2) + R(z)$$

with

$$R(z) = e^{-z} - e^{-z}e_{d-1}(\tfrac{z}{2})F(z/2).$$

Furthermore,

$$R^*(0)/L = -\frac{1}{L}\sum_{j\ge d}\frac{p(j)}{2^j}\Big(\frac{1}{j} + \sum_{i=1}^{d-1}\frac{(j+i-1)!}{2^i i! j!}\Big) - \frac{1}{L}\sum_{i=1}^{d-1}\frac{1}{2^i i!}.$$

For instance, for $d=3$, we get $1 + R^*(0)/L = 0.05432408336\ldots$ which checks well with the values $p(n)$ that we computed up to $n = 500$.

References

[1] J.-P. Allouche and H. Cohen. Dirichlet series and curious infinite products. *Bulletin of the London Mathematical Society*, 17(6):531–538, 1985.

[2] J.-P. Allouche, H. Cohen, M. Mendès France, and J. O. Shallit. De nouveaux curieux produits infinis. *Acta Arithmetica*, 49(2):141–153, 1987.

[3] J.-P. Allouche and J. Shallit. *Automatic sequences.* Cambridge University Press, Cambridge, 2003. Theory, applications, generalizations.

[4] A.D. Barbour, L. Holst, and S. Janson. *Poisson Approximation.* Oxford University Press, 1992.

[5] L.V. Bogachev, A.V. Gnedin, and Y.V. Yakubovich. On the variance of the number of occupied boxes. 2006. available at arXiv:math.Pr/0609498 v1.

[6] P. Flajolet, X. Gourdon, and P. Dumas. Mellin transforms and asymptotics: Harmonic sums. *Theoretical Computer Science*, 144:3–58, 1995.

[7] P. Flajolet and G.N. Martin. Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences*, 31:182–209, 1985.

[8] P. Flajolet and B. Sedgewick. *Analytic combinatorics.* To appear, 2006. available at http://algo.inria.fr/flajolet/Publications/.

[9] W.M.Y. Goh and P. Hitczenko. Gaps in samples of geometric random variables. 2006. available at http://www.math.drexel.edu/∼phitczen/papers.html.

[10] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics (Second Edition).* Addison Wesley, 1994.

[11] P. Hitczenko and A. Knopfmacher. Gap-free samples of geometric random variables. *Discrete Mathematics*, 294:225–239, 2005.

[12] P. Hitczenko and G. Louchard. Distinctness of compositions of an integer: a probabilistic analysis. *Random Structures and Algorithms*, 19(3,4):407–437, 2001.

[13] H.K. Hwang and S. Janson. Local limit theorems for finite and infinite urn models. 2006. available at arXiv:math.PR/0604397.

[14] P. Jacquet and W. Szpankowski. Analytic depoissonization and its applications. *Theoretical Computer Science*, 201:1–62, 1998.

[15] M. Loève. *Probability Theory.* D. Van Nostrand, 1963.

[16] G. Louchard. Probabilistic analysis of adaptative sampling. *Random Structures and Algorithms*, 10:157–168, 1997.

[17] G. Louchard and H. Prodinger. Asymptotics of the moments of extreme-value related distribution functions. *Algorithmica*, 46:431–467, 2006. Long version: http://www.ulb.ac.be/di/mcs/louchard/moml.ps.

[18] G. Louchard and H. Prodinger. On gaps and unoccupied urns in sequences of geometrically distributed random variables (long version). 2006. http://www.ulb.ac.be/di/mcs/louchard/gaps11.ps.

[19] G. Louchard, H. Prodinger, and M.D. Ward. The number of distinct values of some multiplicity in sequences of geometrically distributed random variables. *Discrete Mathematics and Theoretical Computer Science*, AD:231–256, 2005. 2005 International Conference on Analysis of Algorithms.

[20] H. Prodinger. Compositions and Patricia tries: no fluctuations in the variance! *Proceedings of the 6th Workshop on ALENEX and the 1st workshop on ANALCO, L. Arge, G. Italiano, R. Sedgewick (editors)*, pages 211–215, 2004.

[21] H. Prodinger. Periodic oscillations in the analysis of algorithms—and their cancellations. *Journal of the Iranian Statistical Society*, 3:251–270, 2004.

[22] J. O. Shallit. On infinite products associated with sums of digits. *Journal of Number Theory*, 21(2):128–134, 1095.

[23] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences.* Wiley, New York, 2001.